CS8803 BDS / CS4365

Homework Assignment 4

(Programming Category)

Student Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Student Session: cs8803 or CS4365 (circle one)

In this 4th programming assignment, you are given two types of programming problems to gain hand-on experience with ensemble learning. You are required to choose only one of the two problems.

**Post Date**: Monday of Week 9 (Oct. 19)

**Due Date**: Midnight on Saturday of Week 11 (Oct. 31 with no penalty grace period until midnight on Sunday of Nov. 1)

**Problem 1. Hand-on Experience with** **Clustering Ensemble**

Your task for this assignment is to select 4~5 clustering algorithms, such as different implementations of k-means, or different clustering algorithms, and run all of them on a chosen dataset and then use a clustering ensemble method, such as majority voting, to show how much clustering ensemble can improve the learning accuracy. Concretely,

1. Select a dataset from the UCI repository (<http://archive.ics.uci.edu/ml/datasets.html>) or use a dataset of your own choice.
2. Determine how you will measure the quality of the clusters produced. Some reference on clustering evaluation can be found at <http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html>.
3. Select 4-5 algorithms (e.g., different implementations of K-means, canpopy) and run each of them on the same dataset and compare their results using your quality metrics. Note that by varying the initial points, you can obtain different implementations of a k-means algorithm.
4. Design a clustering ensemble algorithm, such as majority voting, to show whether and how much clustering ensemble can improve the learning accuracy.
5. Write a brief report to:

* Describe the dataset and your quality metrics.
* Describe your experiment setup such as how you preprocessed the data (if any), how you chose the parameters for the selected algorithms, and why.
* Present the experiment results for all six methods (4~5 individual clustering and 1 clustering ensemble) in a tabular or chart format for easy comparison.
* Discuss the insights and conclusions from your experiments.  For example, do different clustering methods make a difference in terms of quality or performance for the particular dataset you selected?  Why does clustering ensemble can improve the learning quality?

6. Deliverable.

* One tar or zip file that contains your source files (if available), the executable, a readme file explaining how to compile/run your program.
* The output file for the test dataset screen shots of your execution process.
* Runtime statistics in excel plots or tabular format.
* Report in ppt/word/pdf.

**Problem 2. Hand-on Experience with** **Classification Ensemble**

Your task for this assignment is to select 4~5 classification algorithms and run all of them on a chosen dataset and then use a classification ensemble method, such as majority voting, to show how much an ensemble method can improve the learning accuracy.

1. Select a dataset from the UCI repository (<http://archive.ics.uci.edu/ml/datasets.html>) or use a dataset of your own choice.
2. Determine how you will measure the quality of the clusters produced. Some reference on clustering evaluation can be found at <http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html>.
3. Choose 4~5 different implementations of one classifier, such as the C4.5 decision tree classifier or Naïve Bayesian classifier or SVM or any other classifier that you are familiar with. You can find them from Weka (<http://www.cs.waikato.ac.nz/ml/weka/>), Mahout (<http://mahout.apache.org/users/classification/>) or R (<http://www.rdatamining.com>).
4. Evaluate the individual classifiers using the chosen dataset from UCI repository. One example is the [mushroom dataset](http://archive.ics.uci.edu/ml/datasets/Mushroom) from UCI repository. The training dataset contains 7423 records and the test dataset 701 records.  The first attribute is the class of each record and the rest 21 attributes are categorical attributes.
5. Write a brief report that include the following: Present and discuss the results of your experiments on the chosen dataset with each of the chosen individual classifiers and the classification ensemble; and discuss the experiences and lessons you have learned from the experimentation.
6. Deliverable:

* One tar or zip file that contains your source files (if available), the executable, a readme file explaining how to compile/run your program.
* The output file for the test dataset screen shots of your execution process.
* Runtime statistics in excel plots or tabular format.
* Report in pdf/word/ppt.